dtSearch® and Rosette® Full-Featured International Search

Table of Contents

Synopsis

Search had gone from a convenience to an indispensable requirement of doing business. In enterprises, governments, and non-profits, software engineers are tasked with finding a search engine which is both full-featured and which can quickly retrieve relevant search results in English or any language worldwide. This whitepaper goes into technical detail about the why's and how's of one widely used solution to that problem: the combination of the dtSearch Engine integrated with the linguistic capabilities of Basis Technology's Rosette.

Overview, dtSearch Engine

dtSearch products can instantly search terabytes of text across a desktop, network, Internet or Intranet site. Using the dtSearch Engine, developers can embed dtSearch indexing, searching and file parsing support into their own applications. The dtSearch Engine offers native 64-bit and 32-bit Windows and Linux APIs for C++, Java and .NET.

The dtSearch Engine features over 25 full-text and fielded data search options. Search features cover support for all Unicode-based languages (including Boolean searching, proximity searching, fuzzy searching, etc.). Search features also cover federated searching, unlimited web-based concurrent searching, special forensics search options, and advanced data classification objects.

Using dtSearch's own file parsers and converters, the dtSearch Engine can highlight hits in a wide range of data. The parsers and converters cover MS Office (Word, Excel, PowerPoint, Access), OpenOffice, ZIP, HTML, XML, PDF and many other common file types. dtSearch products also support Exchange, Outlook, Thunderbird and other popular email formats, including attachments.

The built-in Spider works with static and dynamic web data. The Spider covers local and remote, public and private web site data. The Spider is accessible to programmers through a .NET API. The dtSearch Engine also includes APIs for SQL-type data, including BLOB data.

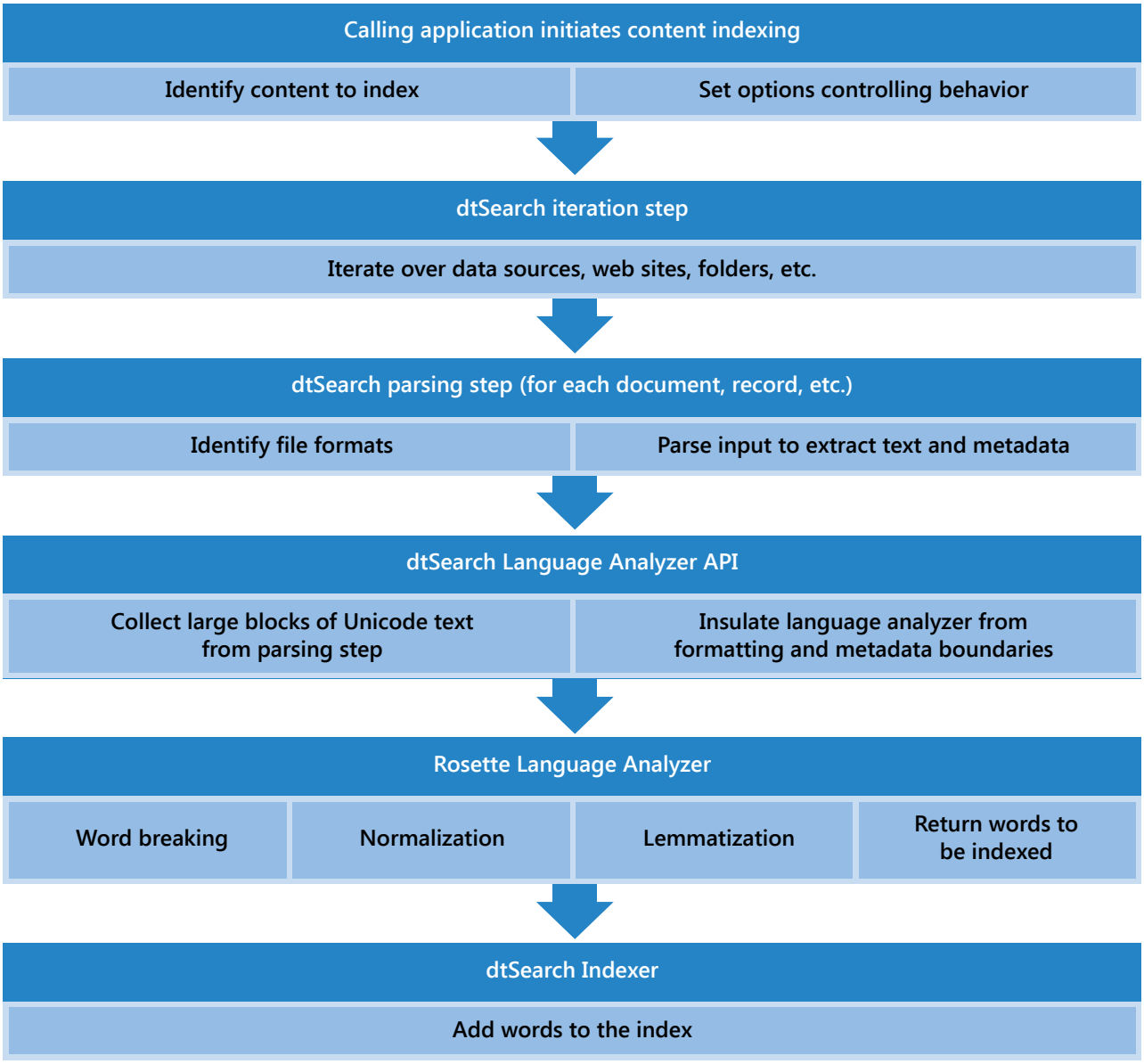
Please visit www.dtsearch.com for hundreds of reviews and case studies, and fully-functional evaluation versions.

dtSearch & Rosette Integration

The dtSearch Engine provides a wide range of search options across a broad spectrum of different data types. While the dtSearch Engine can search Unicode-based languages, the dtSearch Engine does not, however, provide international language morphological analysis capabilities. This section describes how the Rosette API can provide this capability to dtSearch Engine users.

Rosette's functionality integrates with the dtSearch Engine using the dtSearch Engine's Language Analyzer API. Essentially, the API passes blocks of Unicode text to the Rosette language analyzer and accepts back words to index.

The dtSearch Engine / Rosette integrating API is available for evaluation through Basis Technology. More details on the benefits of integration with Rosette follow below.



Benefits of Rosette Integration for International Language Searching

The concepts of precision and recall measure how well search is able to find what the user wants. More precise results with greater recall mean better quality search, but frequently increasing one means a decrease in the other, so the trick is how to maximize each metric while minimizing the impact on the other. A key ingredient in "good search" is an engine that can understand and handle the idiosyncrasies of the language being searched to bring forth the best results.

Let's look at the language-specific processing required to increase recall and/or precision for three categories of notoriously difficult language groups to process: Chinese/Japanese/Korean, Germanic/Scandinavian, and Arabic.

Improving Precision in Chinese, Japanese, and Korean

Asian languages like Chinese, Japanese, and Korean are fundamentally more difficult to process than European languages because their words are not consistently separated by spaces. Chinese and Japanese use no spaces between words, and Korean uses some spaces, but not between every word, and inconsistently depending on the writer. N-gram and Tokenization are methods used to produce indexable units in these languages, each resulting in different degrees of search precision.

N-gram vs. Tokenization

Some engines may use the n-gram technique to break up streams of text into overlapping 2-4 character units, but though recall will be high, precision will be low and indexes will be very, very large, impacting speed.

Here is a Japanese example: 東京都の観光地 (translation: sightseeing spots in Tokyo)

Bigram	English Translation
東京	Tokyo
京都	Kyoto
都の	Capital's
の観	*Not a word*
観光	Sightseeing
光地	*Not a word*

A seven character phrase produces six items to index via the n-gram technique, but also introduces a false word "Kyoto." Since many of the bigrams are not actual words in the original text, many other unrelated strings might accidentally match them, too, and a search becomes a statistical process: Do enough of these two-character segments appear in a given document to flag it as a "match," while not accidentally matching the wrong strings?

This approach also produces many more entries in the index than words, unnecessarily increasing its size. Processing a 100-character buffer (about 13-17 words), adds 999 entries to the index.

A better choice is true tokenization based on morphological analysis,[1] which breaks up text into real words. With tokenization of the example above, there are at most three items to index, and possibly only two if the possessive marker is treated as a stop word.[2]

[1] Morphological analysis in linguistics examines the smallest units of meaning within a word — called "morphemes — such as "un-" in "unremarkable."

[2] A stopword is one which occurs so frequently within a language as to not help in finding search results. Some search engines may choose to ignore them.

Bigram	English Translation
東京都	Tokyo
の	*(possessive marker)*
観光地	Sightseeing

Character Normalization to Improve Recall

Chinese, Japanese, and Korean have additional character normalization needs beyond making sure ASCII characters are all uppercase or lowercase. In digital files, these languages also use a full-width variant of ASCII letters and punctuation which need to be normalized to the half-width ASCII form.

ＡＳＣＩＩｗｏｒｄｓ％＄＆ → ASCIIwords%$&

In the case of Japanese, a half-width version of Japanese *katakana* characters also needs to be normalized to the usual full-width *katakana*. The half-width version was invented in the early days of computer processing to reduce data size, but it is still found in documents and webpages.

ｶﾀｶﾅ → カタカナ

Improving Recall with Pan-Chinese Search

Chinese search gains an enormous boost in recall if both the simplified and traditional Chinese scripts are searched. Mainland China and Singapore use simplified Chinese, whereas Taiwan and Hong Kong use traditional Chinese. Chinese speakers expect one query in one script to find results from *both scripts*.

Pan-Chinese search — i.e., searching documents in both Chinese scripts — is accomplished by converting all the text of one script into the other at query and index time.

The differences between the two scripts fall into three categories.

Category	Simplified Chinese	Traditional Chinese	English Translation
Same character used in both scripts — this is true where the character was simple to start with	大	大	big
Two characters with same pronunciation (but different base meaning) — collapsed to one character in simplified Chinese	头发 出发	頭髮 出發	hair emit
Different vocabulary (analogous to American and British English differences such as “truck” vs. “lorry”) — occurs in mostly modern words	计算机	電腦	computer

In the first case, the change is trivial, and amounts to making sure all the text is in the same encoding;[3] however, the second and third categories require dictionary data to ensure that the conversion is context-sensitive, so that the correct traditional character is chosen.

[3] Generally speaking, when legacy, non-Unicode encodings are used traditional Chinese is written in the Big5 encoding whereas simplified Chinese is encoded in GB2312.

Improving Recall in Germanic and Scandinavian Languages

Linguistic processing required by German, Dutch, Finnish, and Scandinavian languages (Danish, Norwegian, Swedish) is very similar in that these languages freely use compound words which frequently need to be broken up to increase recall.

Take these two German compound words, for example:

German Compound	German Decompounded	English Translation
Jugendarbeitslosigkeit	*Jugend + arbeitslosigkeit*	Youth unemployment (Youth + unemployment)
Samstagmorgen	*Samstag + morgen*	Saturday morning (Saturday + morning)

It is reasonable to imagine that a person looking for information about youth unemployment in German would also welcome search results which included "youth" and "unemployment" as separate words. Similarly, the concepts of "Saturday" and "morning" are very likely to appear as a compound word or as separate words in related documents.

Decompounding increases search recall with minimal impact on precision, but again, requires dictionary data to do properly.

Additionally, to maximize recall, character normalization is needed for words such as "Garten" (garden, singular form) and its plural form "Gärten."

Improving Recall in Arabic

Arabic is one language which especially suffers from low recall if a search engine does not perform Arabic specific linguistic processing, starting from the basic character normalization up to stemming of proper nouns and lemmatization of common nouns. Arabic attaches so many affixes (often prepositions, numbers, and conjunctions) to the beginning, end, and middle of words that without lemmatizing (finding the dictionary word form) before searching, many relevant results are never found.

Character Normalization to Improve Precision and Recall

In English, search engines perform some basic normalization such as lowercasing all the words. In Arabic, normalization is much more complex and will improve both precision and recall. Types of character normalization required include:

- Words with additional vocalization marks such as: سِياسِي vs. سياسي
- Words containing certain letters with dots added or removed such as: قـريـه vs. قـريـة
- Words – including ambiguous cases – containing certain letters with symbols added or removed such as:

  أعتماد vs. اعتماد

  داؤود vs. داوود

  اثم vs. (آثم or إثم or أثم)

Improving Recall with Lemmatization and Stemming

In Arabic, as other languages, lemmatization increases the recall of search results, but with a twist. In Arabic, lemmatization is only applicable to verbs and common nouns (e.g., apple, book, table) *and is not applicable to proper nouns* (e.g., names of people such as, Abul-qassem El-Chabby or Baqah Al-Sharqiyyah).

Proper nouns require stemming to remove prepositions and conjunctions which are often attached to them. Searches for names of people, places, and organizations are seriously hampered without stemming. For example, these phrases in English appear as one word in Arabic: "for Othman," "with Othman," "as Othman," and "and Othman." Thus a search for just "Othman" would not find the above variations without stemming.

Additionally, since Arabic names are frequently the same words as common nouns, part-of-speech tagging becomes critical to differentiating between nouns and proper nouns, particularly when a word's part-of-speech varies depending on where it appears in a sentence.

The Rosette Option for Language Support

Rosette is an SDK designed for a wide range of large-scale applications that need to identify, classify, analyze, index, and search unstructured text from various sources. Rosette's linguistic analysis capabilities are widely used by search engines to increase both precision and recall. It uses a combination of statistical models, dictionary data, and sophisticated computational linguistics to parse digital text in English and over 40 major European, Asian, and Middle Eastern languages.

Years of development and linguistic analysis have gone into the development of Rosette to satisfy the most demanding customers, both for quality, robustness, and speed. In fact, nearly every major web search engine since 1999 — including Bing, goo (Japanese), Google, and Yahoo! — has used Rosette. New generations of search-based applications for e-discovery, financial compliance, and other fields also use Rosette.

Rosette is a cross-platform SDK available for Windows and Unix, and offers all its capabilities via a single API, in C, C++, Java, or .NET.

Rosette Capabilities[4]

- Language Identification – identification of the primary language of a document — or language regions within a multilingual document — and the file's encoding in 55 languages and 45 encodings.
- Unicode Conversion – converts documents in legacy encodings to Unicode
- Character Normalization – normalizes characters to a single representation (e.g. character+ diacritic to single character with diacritics, half/full-size variants of ASCII characters in Asian languages, and several language-specific normalizations.)
- Linguistic Analysis — lemmatization, tokenization, part-of-speech tagging, decompounding, and more in over 40 languages.
- Entity Extraction – extracts entities such as people, places, and organizations in 15 languages via statistical models with customizable user-defined entities via regular expressions and entity databases
- Name Matching – returns matching names despite spelling variations, initials, nicknames, missing name components, missing spaces, out-of-order name components, the same name written in different languages, and more — supported in multiple languages including Arabic, Chinese, English, Korean, and Persian.
- Name Translation – translates names from non-Latin script languages to English and standardizes already translated names — supported in multiple languages including Arabic, Chinese, English, Korean, Russian, and Persian.

The integrated functions of Rosette connect smoothly into the dtSearch Engine to provide linguistic intelligence for many commercial and government applications in a convenient, comprehensive package.

Specifically, Rosette provides language support to enhance the existing international capabilities of the dtSearch Engine, sharpening precision and recall in the ways discussed earlier.

[4] Capabilities are as of Rosette version 7.4.

Chinese, Japanese, and Korean Search

Normalization of Chinese, Japanese, and Korean characters is offered by Rosette's Core Library for Unicode:

- For Chinese, Japanese, and Korean: Converting full-width ASCII characters (Ａ Ｓ Ｄ Ｆ ＆ ％ ＄) to the usual half-width characters (ASDF&%$).
- For Japanese: Converting half-width Japanese *katakana* characters (ｱｲｳｴｵ; one-byte characters invented in the early days of Japanese computing to save on storage) to the usual full-width characters (アイウエオ).

For users who need high quality tokenization, adding Rosette to the dtSearch Engine enables complete analysis of each word's structural components keeping the index slim and the search precision high.

Pan-Chinese Search

Rosette's Chinese script converter function can be called by the dtSearch Engine at index and query time to convert all text into either simplified or traditional Chinese.

Germanic and Scandinavian Languages

Rosette's decompounding for German, Dutch, Danish, Norwegian, and Swedish is backed by dictionary data and easily added to the dtSearch Engine.

Arabic

Adding Rosette to the dtSearch Engine enables character normalization as well as part-of-speech tagging and lemmatization for Arabic.

Comprehensive Name Search

Rosette will match name variations including those due to nicknames, initials, missing name components, missing spaces between names, out-of-order name components, the same name represented in different languages, and more. Thus, besides handling "Robert" vs. "Bob," and "Abdul Rasheed" vs. "Abd-al-Rasheed" vs. "Abd Ar-Rashid," Rosette also matches "Mao Tse-Tung," "Mao Zedong," and "毛泽东" (simplified Chinese).

Summing Up

The chart below summarizes the discussion above and shows how Rosette enhances language support.

Issue Addressed	Enhanced Support from Rosette	Impact on Search with Rosette
Character normalization	*Asian character normalization:* Full-width ASCII characters normalized to half-width for Chinese, Japanese, and Korean. Half-width Japanese *katakana* characters normalized to full-width.	Using Rosette increases recall in Chinese, Japanese, and Korean
Improving Arabic recall by accounting for various word forms or conjugations	Lemmatization for context-sensitive lemmatization and stemming when necessary.	Using Rosette increases recall with minimal loss of precision.
Decompounding of Germanic and Scandinavian languages	Dictionary-based decompounding	Using Rosette increases recall with minimal loss in precision.
Tokenization of languages without spaces between words (Chinese, Japanese, Korean)	Complete morphological analysis and word segmentation for Chinese, Japanese, and Korean.	Using Rosette increases precision and shrinks indexes.

Reliability and Support

Even in a technologically sophisticated company, the core business logic will be the main focus of all business operations, so technical support as a safety net is essential. Just as firms rely on calling the vendor when the copier is acting up, a few judicious pieces of advice from the search or linguistics experts to settle a nettlesome problem lends assurance to both engineers and upper management.

Since 1991, the dtSearch Engine has been providing this high level of assurance for its search engine, which is used worldwide in organizations small and large, spanning almost every business domain.

On the language support side, the wide language coverage of Basis Technology's Rosette — over 40 languages covering Asia, Europe, and the Middle East — ensures there will be one point of contact to address questions about any language, instead of a multitude of single-language vendors with varying support contracts. Rosette developers are on the front lines, providing high quality support to customer inquiries.

Additionally, as a single platform, Rosette's benchmarks for speed and accuracy are readily available, and implementing one or over 40 languages is the same amount of work.

Summary & Conclusion: Rosette and dtSearch, a Powerful Combination

The dtSearch Engine is a full-featured solution that is continually expanding its feature set to match the needs of its customers. The dtSearch Engine combines all the features demanded by developers of high-end commercial applications with a robust, high-performing code base to meet the current and future performance challenges of dealing with ever-larger data sets.

And, for customers who need enhanced search quality for multiple, foreign languages, the Rosette platform fills that gap through solid linguistic analysis technology and comprehensive dictionary data to perform:

- Language Identification
- Linguistic Analysis – Rosette performs a complete linguistic analysis to improve search recall and precision for some of the most linguistically difficult languages. Results integrate easily into dtSearch with minimal performance impact.
  - Lemmatization and Stemming – for boosting recall while maintaining precision in Arabic
  - Tokenization – for Chinese, Japanese, and Korean which do not have spaces between words. Although dtSearch users have access to a tokenizer, Rosette provides yet deeper morphological analysis of the words.
  - Chinese script conversion – for pan-Chinese search
  - Character normalization – specialized transforms for Arabic and Asian languages
  - Decompounding – for Germanic and Scandinavian languages which form words from multiple words
  - Part-of-speech tagging – for context-sensitive lemmatization, and particularly for Arabic search to distinguish between common nouns requiring lemmatization, and proper nouns, requiring stemming
- Entity extraction – To locate names, places, organizations, and other entities, both to boost ranking of results whose entities match the search query or as a base for building faceted search.
- Name matching – Rosette can match names that differ due to nicknames, initials, missing name components, missing spaces between names, out-of-order name components, the same name represented in different languages, and more.

Next Steps

- To learn more about dtSearch, or to try a fully-functional dtSearch Engine evaluation version, please visit www.dtsearch.com
- To request a free product evaluation of Rosette, or to request an evaluation copy of the dtSearch Engine/ Rosette integrating API, please contact Basis Technology at: http://www.basistech.com/products/requests/evaluation-request.html

About Basis Technology

Basis Technology (www.basistech.com) provides software solutions for text analytics, information retrieval, and name resolution in many languages. The Rosette® linguistics platform is a widely adopted suite of interoperable components that delivers high-performance results to search, business intelligence, e-discovery, and many other enterprise applications. Basis Technology is on the forefront of applied natural language processing solutions using a combination of statistical modeling, expert rules and corpus-derived data.

Leading software vendors, content providers, financial institutions, and government agencies rely on Basis Technology's solutions for Unicode compliance, language identification, multilingual search, normalization, name matching, name translation, and entity extraction. Our products and services are used by over 250 major firms, including Cisco, EMC, Endeca, Exalead/Dassault, Fujitsu, Hitachi, HP, Microsoft, Oracle, and Software AG. Our text analysis products are widely used in the U.S. defense and intelligence industry by such firms as CACI, Lockheed Martin, Northrop Grumman, SAIC, and SRI. We are also the top provider of multilingual search technology to web search engines, such as AOL, Ask.com, Google, Microsoft/Bing, and Yahoo!

Company headquarters are in Cambridge, MA, with branch offices in San Francisco, California; Herndon, Virginia; and Tokyo, Japan. For more information, visit www.basistech.com or call 1-800-697-2062.

About dtSearch

The Smart Choice for Text Retrieval® since 1991, dtSearch offers 20+ years of experience in parsing and searching data. The dtSearch product line includes enterprise and developer text search products, meeting some of the largest-capacity text retrieval needs in the world. dtSearch products have received hundreds of excellent case studies and press reviews. (Please see www.dtsearch.com for these.) The company has distributors worldwide, including coverage on six continents.

For more information, or to obtain fully-functional evaluation versions, please visit www.dtsearch.com, call 1-800-IT-FINDS (1-800-483-4637), or email sales@dtsearch.com.